Unified Structure and Content Search for Personal Information Management Systems

Wei Wang, Amélie Marian, Thu D. Nguyen
{*ww, amelie, tdnguyen*}*@cs.rutgers.edu*

Technical Report DCS-TR-661
Department of Computer Science, Rutgers University
110 Frelinghuysen Rd, Piscataway, NJ 08854

December 14, 2009

Abstract

The amount of data that users are storing and accessing in personal information systems is growing massively. At the same time, the organization of this data is becoming more heterogeneous, with data spread across different organizational domains such as emails, music databases, and photo albums, some of which are structured by applications rather than users. Powerful search tools are needed to help users locate data in these rapidly expanding yet fragmented data sets. In this paper, we present a novel fuzzy search approach that considers approximate matches to structure and content query conditions. Our approach includes a scoring framework for computing unified relevance scores for potential answers. Critically, our framework uses unified data and query processing models so that structure conditions can be approximately matched by content inside files and vice versa. Our model also unifies external structure (directories) with internal structure (e.g., XML structure), allowing users to specify integrated queries that are matched to a single unified data domain. We propose indexes and algorithms for efficient query processing. Finally, we empirically evaluate our approach using a real data set. We show that our unified fuzzy search approach can leverage structure information to significantly improve search accuracy, yet is robust to mistakes in query conditions.